PB80-178379

# Evaluation of the Real-Time Air-Quality Model Using the RAPS (Regional Air Pollution Study) Data Base. Volume 4. Evaluation Guide

SRI International
Menlo Park, CA

Prepared for

Environmental Sciences Research Lab.
Research Triangle Park, NC

Feb 80

U.S. Department of Commerce
National Technical Information Service
NTIS

PB80-178379

United States
Environmental Protection
Agency

Environmental Sciences Research
Laboratory
Research Triangle Park NC 27711

EPA-600/4-80-013d
February 1980

Research and Development

EPA

# Evaluation of the Real-Time Air-Quality Model Using the RAPS Data Base

## Volume 4. Evaluation Guide

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

## RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into nine series. These nine broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The nine series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies
6. Scientific and Technical Assessment Reports (STAR)
7. Interagency Energy-Environment Research and Development
8. "Special" Reports
9. Miscellaneous Reports

This report has been assigned to the ENVIRONMENTAL MONITORING series. This series describes research conducted to develop new or improved methods and instrumentation for the identification and quantification of environmental pollutants at the lowest conceivably significant concentrations. It also includes studies to determine the ambient concentrations of pollutants in the environment and/or the variance of pollutants as a function of time or meteorological factors.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

**TECHNICAL REPORT DATA**
*(Please read Instructions on the reverse before completing)*

| | |
|---|---|
| 1. REPORT NO. EPA-600/4-80-013d | 2. |
| 3. RECIPIENT'S ACCESSION NO. PB 80 178379 | |
| 4. TITLE AND SUBTITLE<br>EVALUATION OF THE REAL-TIME AIR-QUALITY MODEL USING THE RAPS DATA BASE<br>Volume 4. Evaluation Guide | 5. REPORT DATE<br>February 1980 |
| | 6. PERFORMING ORGANIZATION CODE |
| 7. AUTHOR(S)<br>Ronald E. Ruff | 8. PERFORMING ORGANIZATION REPORT NO.<br>Final Report<br>SRI Project 6868 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>SRI International<br>333 Ravenswood Avenue<br>Menlo Park, California 94025 | 10. PROGRAM ELEMENT NO.<br>1AA603 AA-26 (FY-77)<br>11. CONTRACT/GRANT NO.<br>68-02-2770 |
| 12. SPONSORING AGENCY NAME AND ADDRESS<br>Environmental Sciences Research Laboratory — RTP, NC<br>Office of Research and Development<br>U.S. Environmental Protection Agency<br>Research Triangle Park, North Carolina 27711 | 13. TYPE OF REPORT AND PERIOD COVERED<br>FINAL 8/77-4/79<br>14. SPONSORING AGENCY CODE<br>EPA/600/09 |
| 15. SUPPLEMENTARY NOTES | |

16. ABSTRACT

The theory and programming of statistical tests for evaluating the Real-Time Air-Quality Model (RAM) using the Regional Air Pollution Study (RAPS) data base are fully documented in four volumes. Moreover, the tests are generally applicable to other model evaluation problems. Volume 4 discusses the application and interpretation of the statistical programs, particularly with regard to use on the RAM. In general, there is no set procedure for evaluating an air-quality model because of the different reasons for evaluating models and many subjective decisions to be made during the process. However, guidelines are presented to cover a wide variety of evaluation needs, with attention to data preparation, classification, analysis, selection and application of tests, and interpretation of results. Several methods of diagnosing causes of poor model performance are discussed and some sample program outputs are also provided.

17. KEY WORDS AND DOCUMENT ANALYSIS

| a. DESCRIPTORS | b. IDENTIFIERS/OPEN ENDED TERMS | c. COSATI Field/Group |
|---|---|---|
| * Air Pollution<br>* Mathematical models<br>* Evaluation<br>* Tests<br>* Computer systems programs<br>* Statistical tests | Real-Time Air-Quality Model<br>Regional Air Pollution Study<br>Data Base | 13B<br>12A<br>14B<br>09B |

| | | |
|---|---|---|
| 18. DISTRIBUTION STATEMENT<br>RELEASE TO PUBLIC | 19. SECURITY CLASS *(This Report)*<br>UNCLASSIFIED | 21. NO. OF PAGES |
| | 20. SECURITY CLASS *(This page)*<br>UNCLASSIFIED | 22. PRICE |

EPA Form 2220-1 (9-73)

EPA-600/4-80-013d
February 1980

# EVALUATION OF THE REAL-TIME AIR-QUALITY MODEL USING THE RAPS DATA BASE

## Volume 4. Evaluation Guide

by

Ronald E. Ruff
Atmospheric Science Center
SRI International
Menlo Park, California 94025

Contract No. 68-02-2770

Project Officer

John S. Irwin
Meteorology and Assessment Division
Environmental Sciences Research Laboratory
Research Triangle Park, North Carolina 27711

ENVIRONMENTAL SCIENCES RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

## DISCLAIMER

This report has been reviewed by the Environmental Sciences Research Laboratory, U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

ABSTRACT

The theory and programming of statistical tests for evaluating the Real-Time Air-Quality Model (RAM) using the Regional Air Pollution Study (RAPS) data base are fully documented in four volumes. Moreover, the tests are generally applicable to other model evaluation problems. Volume 4 discusses the application and interpretation of the statistical programs, particularly with regard to use on the RAM. In general, there is no set procedure for evaluating an air-quality model because of the different reasons for evaluating models and many subjective decisions to be made during the process. However, guidelines are presented to cover a wide variety of evaluation needs, with attention to data preparation, classification, analysis, selection and application of tests, and interpretation of results. Several methods of diagnosing causes of poor model performance are discussed and some sample program outputs are also provided. This report was submitted in fulfillment of Contract No. 68-02-2770 by SRI International under the sponsorship of the U.S. Environmental Protection Agency. This report covers a period from 1 October 1977 to 1 April 1979, and work was completed as of 1 April 1979.

CONTENTS


Abstract . . . . . iii
Figures . . . . . vi
Tables . . . . . vi

1. Overview . . . . . 1
   Introduction . . . . . 1
   General procedures . . . . . 1
2. The Model Evaluation Process . . . . . 3
   Discussion . . . . . 3
   Data preparation . . . . . 4
   Data analysis . . . . . 12
   Selection and application of the statistical tests . . . . . 14
   Interpretation of results . . . . . 16
   Diagnosing the causes of poor model performance . . . . . 25
   Combinations of poor performance indicators . . . . 33

References . . . . . 35
Appendices

A. SPSS multiple regression output for test data base . . . . . 36
B. Partial sample of output for SPSSSORT program . . . . . 45


Preceding page blank

FIGURES


| Number | | Page |
|---|---|---|
| 1 | General procedure for evaluating models | 2 |
| 2 | RAM evaluation data bases | 11 |
| 3 | General logic to compute multihour averages | 13 |
| 4 | Frequency distribution for test data base | 17 |
| 5 | Residual frequency distribution for test data base | 17 |
| 6 | Actual versus ideal fit for test data base | 19 |
| 7 | Scattergram with bounds for the test data base | 19 |
| 8 | Autocorrelation function of test data base | 21 |
| 9 | Normalized cumulative periodogram for test data base | 21 |
| 10 | SPSS summary output for test data base | 24 |
| 11 | Types of bias | 27 |
| 12 | Temporal bias example | 31 |


TABLES


| Number | | Page |
|---|---|---|
| 1 | Wind-Speed Classes | 6 |
| 2 | Mixing-Depth Classes | 6 |
| 3 | Temperature Classes | 7 |
| 4 | Modified Test Data Base Hourly Format | 9 |
| 5 | Description of RAM Evaluation Format and Parameters | 10 |

# SECTION 1

# OVERVIEW

## INTRODUCTION

In the previous volumes, a distinction was made between final-evaluation statistics and intermediate-evaluation statistics. Final-evaluation tests compute accuracy scores, which are used to compare a model's performance against some standard, such as a competing model or a different version of the same model. These accuracy scores do not provide any insight as to why a model is performing poorly (or well). Instead, they provide a bottom-line figure (score) which sums up the performance.

Intermediate evaluation statistics are used to diagnose the reasons for a model's poor performance, either in general or for certain meteorological or emission conditions. In all, there are five tests that can be used for intermediate evaluation: residual time series, chi-square goodness-of-fit, bivariate regression and correlation, interstation error correlation, and multiple regression of error residuals. Graph displays are incorporated into most of the programs to facilitate interpretation.

## GENERAL PROCEDURES

As will be discussed in the following pages, the general method of applying the programs remains essentially the same regardless of the specific air-quality model being tested. To adapt the programs to the RAM evaluation, only minor software modifications must be made. These include: (1) scaling the graphical displays, (2) annotating the output, (3) adjusting DIMENSION statements to accommodate the proper number of observations and stations, and (4) formatting the input data. The documentation on the above items is contained in Volume 3 of this report, "Program User's Guide."

As will be discussed in the following sections, the recommended model evaluation procedure is an iterative process whereby results from one test are used to decide which test to apply next. Before delving into the specifics, consider the generalized procedure shown in Figure 1. Our first step is to analyze the data base in a number of ways. After we find out a little about the data base, we can make a well-founded decision about which statistical test to apply. We then apply that test and analyze the results. Then we can opt to (1) end our evaluation, (2) conduct another test, or (3) modify the model and proceed through the evaluation process again.

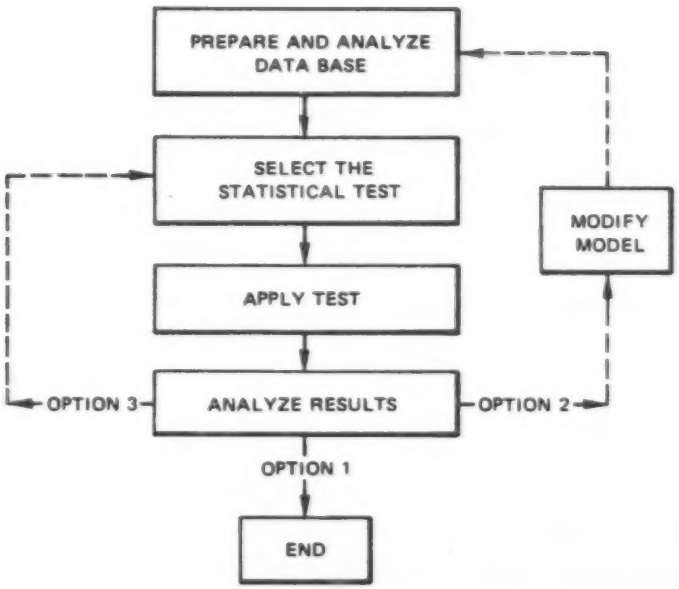



Figure 1. General procedure for evaluating models.

The flow chart in Figure 1 depicts a human-decision process. (It is not a flow chart for a computer program.) Because subjective human decisions are involved, there are many possible variations in the specific choices. Guidelines for these choices are presented in the following sections.

## SECTION 2

## THE MODEL EVALUATION PROCESS

### DISCUSSION

As we discussed in the previous section, there will be many variations possible in our testing strategy for the RAM or any other model. The particular strategy depends on (1) the user-defined evaluation criteria, (2) the interpretation of intermediate test results, and (3) a subjective evaluation of the best course of action. Item (1) above recognizes that the most important evaluation criteria can vary among applications. For instance, one user might be most interested in evaluating performance for a 24-hour averaging period while another may select a 1-hour averaging period. The interpretation of the results of a particular test, item (2), will determine which test will be conducted next and how to stratify the data base accordingly. Finally, for item (3), we recognize that there are many "gray" areas in selecting a future course of action. After evaluation of some intermediate test results, one user may choose to run additional tests while another may decide to modify the model and then rerun the same test.

There are probably numerous reasonable approaches to the model-evaluation process, each of which could produce the same successful results. Therefore, our discussion by necessity must dwell on a general testing procedure, with some assumptions on how the RAM test would proceed. We also attempt to discuss the various alternatives along the way.

For evaluating the RAM, we are assuming that the primary objective is to improve the model for the one-hour averaging time. Secondary objectives would be to compare the RAM with other models; assess its accuracy against some user-defined standards; or evaluate it for other averaging times, such as 3-hour, 24-hour, monthly, seasonal, and annual. For these secondary objectives, the final evaluation statistics are recommended.

As we will be noting in the following sections, some intermediate evaluation statistics may also be helpful, but only to complement the final-evaluation statistics.

In the remainder of this section we discuss the steps in the model-evaluation process, starting with data preparation and continuing through the last computation of the final-evaluation statistics.

## DATA PREPARATION

In this volume we have assumed that the primary objective is to evaluate and improve the RAM by testing it with the RAPS data base. As a start, the test data base consists of hourly values covering a one-year period for (1) observed and predicted concentrations at 13 locations; (2) wind measurements at these locations; (3) derived wind speed and direction, atmospheric stability class, mixing height, and temperature parameters representing the St. Louis area (as a whole); and (4) an emission inventory. The 13 locations are for the monitoring stations in the RAPS network. The observed concentrations are averaged from the $SO_2$ monitoring instruments; the predicted concentrations result from applying the RAM to the input data from (3) and (4) above.

### Classification of Parameters

Except for atmospheric stability all the RAM input data are numerical values which are directly averaged from the instrument readings. To stratify test results according to meteorological and/or emission conditions, the data are divided into classes for each relevant RAM input parameter. The particular parameters and recommended number of classes are:

| Parameter | Number of Classes |
|---|---|
| Atmospheric stability | 7 |
| Wind speed | 6 |
| Wind direction | 16 |
| Mixing height | 7 |
| Temperature | 5 |
| Emissions | 16 |

If we examine the number of combinations from the above classes, we see that there are 376,320 possible categories. Further, we note that there will be only 8,784 hourly $SO_2$ concentrations for the yearly (1976) test data base. Therefore, we will have to combine the parameter categories in a reasonable manner so that within each category there is a statistically valid sample (e.g., 10 cases*). For our purposes, it is better not to limit the number of categories at this time. This is best done after the RAPS data base has been reviewed, as will be discussed later.

Our classification structure, at this point, is taken mostly from the literature. As stated earlier, stability is classified in one of the seven atmospheric stability categories for an urban area, as calculated in the RAM preprocessor. Wind directions are divided into 16 sectors of 22.5° each, with the first sector centered in north.† Wind speed is divided into six categories according to the scheme used by the Climatological Dispersion Model.[2]‡ (See Table 1.) Mixing-depth classes correspond to those used in the APRAC urban diffusion model.[2] (See Table 2.)

Ambient temperature measurements are used both explicitly by RAM, in the plume-rise calculation, and implicitly, in the emissions inventory. ($SO_2$ emissions from space-heating sources are computed when ambient temperatures are below 69°F, or 20°C.) Our temperature categorization scheme uses the space-heating cutoff temperature, 20°C, as a dividing point between the third and fourth categories. Table 3 describes the temperature categories.

For emission categories we suggest using the actual RAPS emission rates. Prior to analyzing the data base, we tentatively suggest dividing the inventory into 4 categories of area-source emissions and 4 categories

---

*For statistical significance, the required number of cases depends on the standard deviation of the sample

†In our terminology, this corresponds to the direction from which the wind is blowing.

‡References are listed at the end of this report.

TABLE 1. WIND-SPEED CLASSES

| Wind-Speed Class | Speed Interval (knots) | Central Wind Speed (m/sec) |
|---|---|---|
| 1 | 0-3 | 1.50 |
| 2 | 4-6 | 2.46 |
| 3 | 7-10 | 4.47 |
| 4 | 11-16 | 6.93 |
| 5 | 17-21 | 9.61 |
| 6 | >21 | 12.51 |

TABLE 2. MIXING-DEPTH CLASSES

| Mixing-Depth Class* | Mixing-Depth Interval (meters) | Geometric Mean of Depth |
|---|---|---|
| 1 | ≤100† | 70.0 |
| 2 | 100-200 | 141 |
| 3 | 200-400 | 283 |
| 4 | 400-800 | 566 |
| 5 | 800-1,600 | 1,131 |
| 6 | 1,600-3,200 | 2,262 |
| 7 | >3,200 | 4,525* |

*For classes 1 and 7, the geometric mean was calculated as though the classes were bounded.

†For the RAPS data base, mixing depths are a minimum of 100 meters.

TABLE 3. TEMPERATURE CLASSES

| Temperature Class | Temperature Interval (°C) |
|---|---|
| 1 | ≤0 |
| 2 | 0-10 |
| 3 | 10-20 |
| 4 | 20-30 |
| 5 | >30 |

of point-source emissions. Alternatively one could use the four seasons of the year, a weekend-weekday separation, and a day-night separation. (This results in 16 possible emissions categories: 4 seasons x 2 weekend-weekday x 2 day-night.) In addition, as mentioned earlier, ambient temperature could also be used in conjunction with the time period to further categorize emissions.

Formats

The format used for our test data base, described in Volume 3, must be modified somewhat for evaluating the RAM. The difference is that for RAM we have derived wind termperature, stability, and mixing-height data, which are used for model input. In addition we have wind measurements averaged hourly for each station, which we also recommend incorporating into the format. (However, this is optional.) Therefore, the recommended data base consists of values averaged hourly for the following:

- Interpolated wind-speed class
- Interpolated wind-direction class
- Mixing-height class
- Temperature class
- Site-specific wind speed (optional)

- Site-specific wind direction (optional)
- RAM-predicted concentration for each site
- RAPS-observed concentration for each site.

For the last four items, there will be 13 observations for each time period. (This corresponds to the number of sites for which $SO_2$ monitoring data are available.) For the first four items there is just one observation per time period. We also suggest adding the following logistic parameters:

- Date (one per time period)
- Time (one per time period)
- Site identification (13 per time period).

Therefore, for a full (leap) year (8,784 hours), the size of our data base is:

$$8{,}784 \times 6 + 8{,}784 \times 5 \times 13 = 623{,}664^{*} \quad .$$

The manner in which these are ordered is considered next.

The general format of the test data base was given in Volume 3. To process the RAM data base efficiently, we suggest a slight modification in the hourly average format as indicated in Table 4. The parameters are defined in Table 5. Note that for each hourly time period there are NS + 1 lines. (We allow that in a mass storage file line numbers are irrelevant. However, it simplifies our discussion here.) In general, the first line of the group contains the categorized equivalent to the RAM input data. The next 13 lines contain the observed and predicted concentrations and the wind measurements for each of the 13 stations.

---

*If we eliminate the site-specific wind speed and direction the size reduces to 394,500 words.

TABLE 4. MODIFIED TEST DATA BASE HOURLY FORMAT

| Line | Data |
|---|---|
| 1 | $D_1, T_1, ASC_1, WSC_1, WDC_1, MDC_1, TP_1, EC_1$ |
| 2 | $1, OC_1, PC_1, WS1_1, WD1_1$ |
| 3 | $2, OC_2, PC_2, WS2_1, WD2_1$ |
| | ⋮ |
| NS + 1 | $NS, OC_{NS}, PC_{NS}, WSNS_1, WDNS_1$ |
| | ⋮ |
| 14 | $13, OC_{13}, PC_{13}, WS13_1, WD13_1$ |
| 15 | $D_2, T_2, ASC_2, WSC_2, WDC_2, MDC_2, TP_2, EC_2$ |
| 16 | $1, OC_{14}, PC_{14}, WS1_2, WD1_2$ |
| 17 | $2, OC_{15}, PC_{15}, WD2_2, WD2_2$ |
| | ⋮ |
| NS + 2 | $NS, OC_{2 \cdot NS}, PC_{2 \cdot NS}, WSNS_2, WDNS_2$ |
| | ⋮ |
| 28 | $13, OC_{26}, PC_{26}, WS13_2, WD13_2$ |
| 29 | $D_3, T_3, ASC_3, WSC_3, WDC_3, MDC_3, TP_3, EC_3$ |

TABLE 5. DESCRIPTION OF RAM EVALUATION FORMAT AND PARAMETERS

| Line/Card Number | Parameter Number | Parameter Name | Format* | Description |
|---|---|---|---|---|
| 1, 15, ... | 1 | D | I6 | Date (yr-mo-day); e.g., 761002 |
| | 2 | T | I4 | Time, 0000-2359 |
| | 3 | ASC | I1 | Atmospheric-stability class, 1-7 |
| | 4 | WSC | I1 | Composite wind-speed class, 1-6 |
| | 5 | WDC | I2 | Composite wind-direction class, 1-16 |
| | 6 | MHC | I1 | Mixing-height class, 1-7 |
| | 7 | TP | I1 | Temperature class, 1-5 |
| | 8 | EC | I6 | Emission class, 1-16 |
| 2-14, 16-28, ... | 1 | NS | I3 | Monitoring station number, 101-125 |
| | 2 | OC | I4 | Observed concentration |
| | 3 | PC | I4 | Predicted concentration |
| | 4 | WSNS | I1 | Site-specific wind-speed class,† 1-6 |
| | 5 | WDNS | I2 | Site-specific wind-direction class,† 1-16 |

*We assume the format will be free-form, with parameters separated by commas. This may not be permitted on some computers.

†May be omitted because they are not directly used in the RAM. However, inclusion is desirable for assessing the adequacy of the composite wind calculations.

For the RAM evaluation, data exist in three forms, as illustrated in Figure 2. These data will be supplied in three segments, as shown on the left-hand side of the figure. The data will be processed into a single data base in the format described in the previous paragraph. Furthermore, we recommend an abbreviated data base consisting only of the observed and predicted concentrations covering the time period of interest for the 13 stations (e.g., for the year). Because most of the statistical tests use only the observed and predicted concentrations, the abbreviated data bases are generated to reduce processing costs. For the year, the maximum size of this test data base is 228,384 words (8,784 hrs x 2 concentrations x 13 sites). However, some of these data bases will be segregated by given meteorological and emissions conditions and thus could be much smaller in size. The SPSS software described in Volume 3 can be used to segregate the data.

## Calculation of Nonhourly Averages

The statistical package can be used to evaluate the RAM for nonhourly averages. Typical averaging periods of interest would include 3-hour, 24-hour, monthly, seasonal, and annual. To form the data base for these periods, a simple FORTRAN program that converts the hourly averaged data into the desired form is recommended.

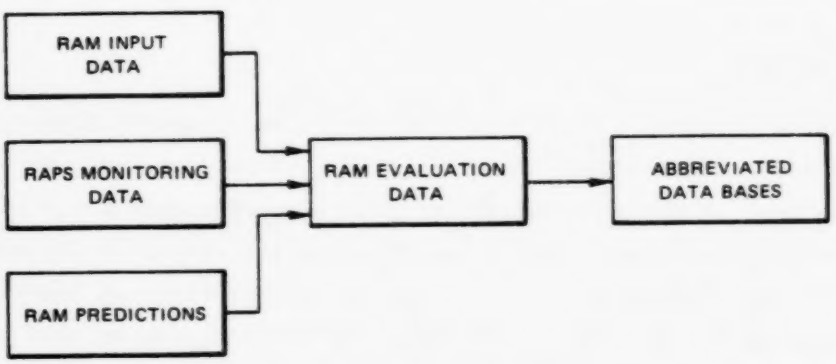



Figure 2. RAM evaluation data bases.

Figure 3 illustrates a suggested logic for the FORTRAN program, which assumes that the data are time-sequential for each group of 13 sites. The date and time are also needed because the time index is computed from the time, for the 3-hour average, and from the date, for the 24-hour, monthly, seasonal, and annual averages. (For instance, if we assume time ranges from 0000 to 2300, the 3-hour average time index can be computed from the following: 1 + ITIME/3. Since the second term will truncate to an integer, we note that the index will range between 1 and 8.) The time index is needed only to indicate when to compute the specified average. After the time index is computed, the program determines if the index has changed. If it has not, the observed (OC) and predicted concentrations (PC) are accumulated, along with the count. If the time index has changed, the observed and predicted concentrations are averaged (if valid) and written on the output file. This process continues until an end-of-file is encountered.

Checks are needed to separate out bad data (denoted by a "-99") so they are not counted in the multihour average. In addition, after the time index changes, the number of data points comprising the average must be checked. Do we need three for the hourly average, or will two suffice? Such criteria must be established for each average of interest. For the RAPS data base, these criteria will be fairly rigid, provided we can get a valid statistical sample.

## DATA ANALYSIS

Prior to conducting statistical tests, we recommend first analyzing the data in an elementary manner to gain a better persepective on the distribution of the observed and predicted concentrations and the categorized RAM input parameters. The EPA program, stored in SRI.FREQ (as documented in Volume 3, p. 13), produces distributions on logarithm-frequency axes. The SPSS[3] procedure, FREQUENCIES, can be used to obtain line-printer histograms for the RAM input parameters (see Volume 3, p. 10).

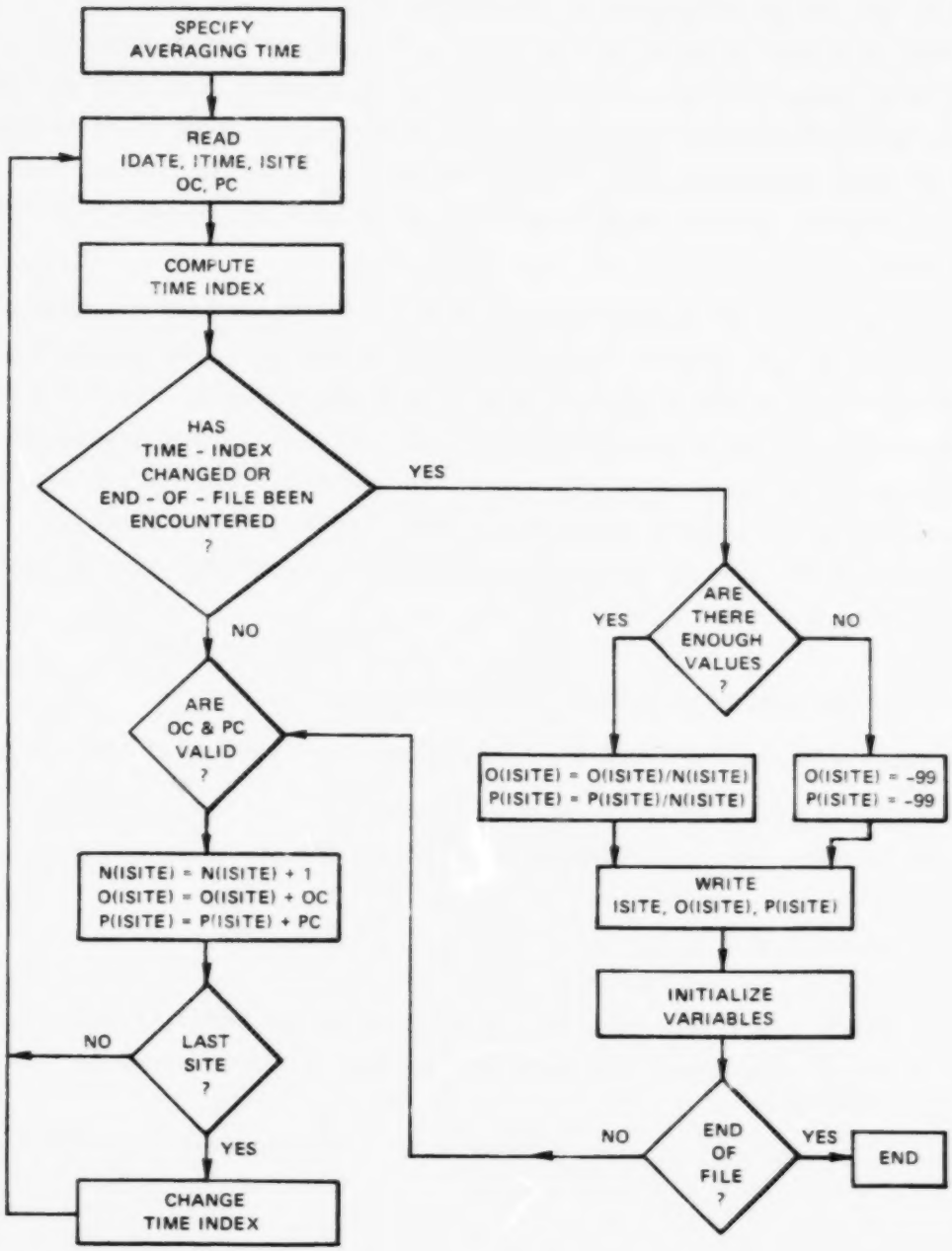



Figure 3. General logic to compute multihour averages.

The distribution of concentrations, observed and predicted, will show whether it is worthwhile to proceed further with out analysis. If the RAPS data base is to prove useful, the distribution must cover a reasonable range of observed concentrations (preferably from zero through levels above the air-quality standards) and a statistically sound sample size of these concentrations. This distribution should be plotted for all 13 sites to provide insight into the usefulness of the data base for each site.

The line-printer histograms of the RAM input parameters indicate how representative our parameter classification scheme is. (The analysis may suggest that a different scheme might be more appropriate. For instance, if 90 percent of the temperatures are contained in two categories we may choose to assign narrower ranges to the categories.) The histograms will also be useful at a later stage in our analysis when we are relating residuals (difference in observed and predicted concentrations) to input-parameter categories.

## SELECTION AND APPLICATION OF THE STATISTICAL TESTS

The basic order of events in evaluating and improving the RAM is

- Apply the final-evaluation statistics
- Run the intermediate-evaluation statistics
- Interpret the results
- Run tests on subgroups
- Modify the model (RAM)
- Apply the final-evaluation statistics to the modified model
- Repeat the above process until the desired improvement is achieved.

In the following subsections, we examine each of the above steps in more detail.

Application of the Intermediate-Evaluation Statistics

The final-evaluation statistics are the gauge by which we judge a model's overall performance. Because our goal is to improve the RAM, its current performance becomes the baseline from which to compare future "improved" versions. Hence, the final-evaluation statistics should be applied for each version of the model.

There are eight unique tests in our final-evaluation package. Because the incremental (computer) cost per test is small compared to the input-output costs, we suggest application of most tests with exceptions as noted below. The following parameter values are suggested (Refer to Appendix B of Volume 2):

- Test 1 - No user-specified parameters required.
- Test 2 - No user-specified parameters required.
- Test 3 - The absolute-error threshold, which is $\epsilon$ = 10 ppb (or 26.1 $\mu g/m^3$); this error threshold is consistent with instrument accuracy determined in the quality-assurance program.[4]
- Test 4 - The percentage-error threshold, which is $\epsilon$ = 0.20. Using the percentage error can be misleading if the data base is biased toward lower concentrations, which is usually the case. Therefore, after analyzing the RAPS data base, we may suggest omitting this test.
- Test 5 - As a start, set the cutoff = 500 ppb (or 1,307.5 $\mu g/m^3$), which corresponds to the Federal Air Quality Standard averaged over 3 hours. This may have to be modified downward if there are not enough concentrations at or above 500 ppb (say, a minimum of 200 hourly averages per station).
- Test 6 - The same criteria (Test 5) for the cutoff value applies here; arbitrarily we suggest a value of 1 for $\ell_1$ and 2 for $\ell_2$. However, this test is very subjective and may not be too useful in the RAM evaluation.
- Test 7 - Again, this is a very subjective test requiring the entry of a 5 x 5 loss matrix. Arbitrarily, we recommend the symmetric loss matrix shown in Appendix B of Volume 2.
- Test 8 - This loss function depends on the distance between the maximum-observed and maximum-predicted locations. Therefore, we recommend a 13 x 13 loss matrix consisting of the distances between the RAPS monitoring stations.

In summary, we suggest conducting the final-evaluation tests for each version of the RAM. For reasons stated above, Tests 4, 6, and 7 should be less valuable than the others.

## Application of the Final Evaluation Statistics

Each of the intermediate evaluation statistics can tell us something different about the RAM's performance. Hence, at least for the first application of intermediate statistics, it is a good idea to run all the tests. In conjunction with the displays described on page 12, the user now has a complete record of statistical figures-of-merit, associated graphic displays and auxiliary displays.

The actual mechanics of running each test are documented in Volume 3. We now focus our attention on how to interpret and categorize the results, prior to conducting more tests and subsequently modifying the RAM.

## INTERPRETATION OF RESULTS

There is some redundancy in our statistics package, which means that conclusions about RAM performance may be arrived at in a number of different manners. When evaluating poor performance, we are looking for one or more types of poor-performance indicators. For each such indicator, there are one or more statistical tests that will tell us something about the RAM performance. These indicators of poor performance include:

- General bias--if the residuals between observed and predicted concentrations are mostly of the same magnitude and sign, then we have a general bias. General bias will be graphically displayed by the following:
  - Logarithm-frequency plots of residuals or observed and predicted concentrations on the same plots. These plots for the test data base are shown in Figures 4 and 5, respectively. From Figure 4 we note, for the 10-percent point, that the difference is less than a factor of 2. Thus, for over 90 percent of the time the agreement between observation and prediction is within a factor of 2. Figure 5 presents another view of the above comparison by displaying the distribution of the residual concentration (magnitude and sign).

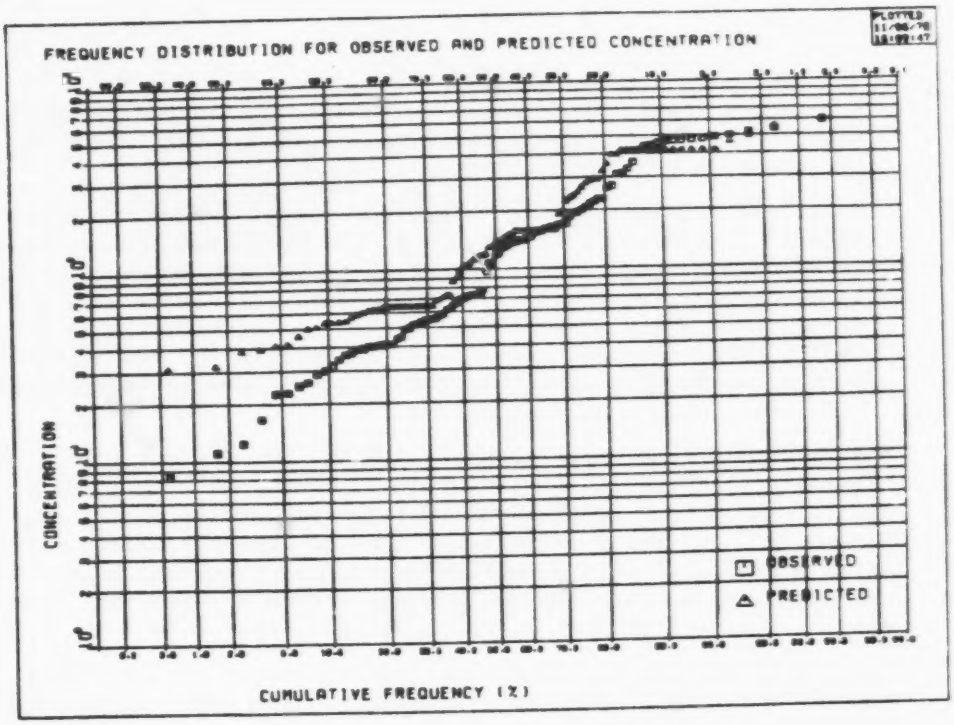



Figure 4. Frequency distribution for test data base.

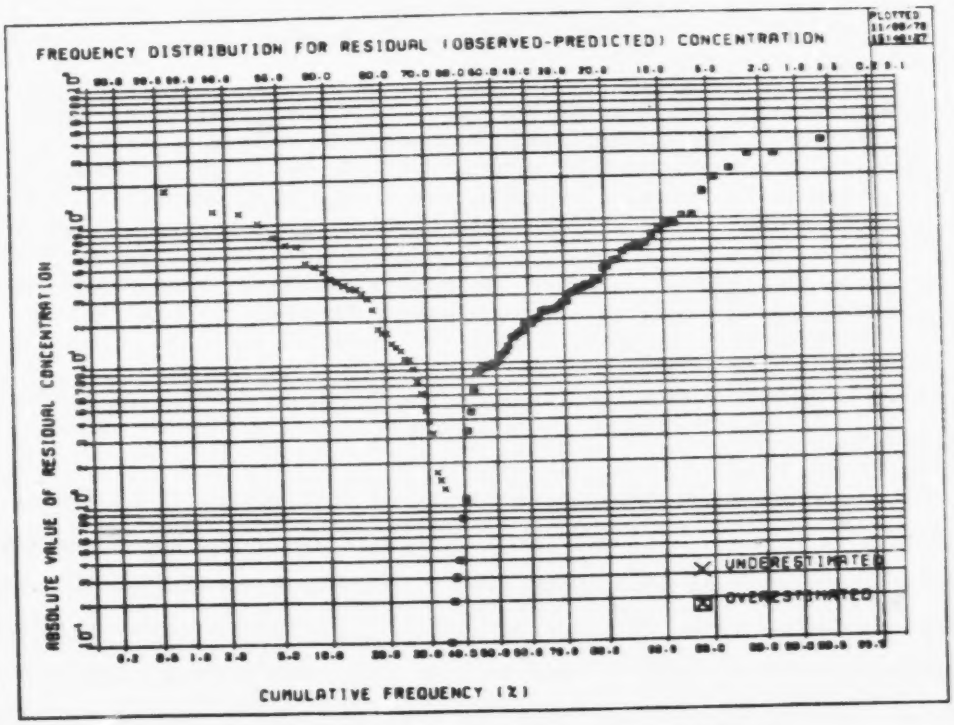



Figure 5. Residual frequency distribution for test data base.

- Chi-square test. The graphic output of this test closely parallels that of the log-frequency plots above, but is not as effective in displaying bias.
- Bivariate regression and correlation test. Two of the plots resulting from application to the test data base are presented in Figures 6 and 7. In Figure 6, a comparison between the least-squares regression line and "ideal" agreement line provides perhaps the most valuable indicator of bias. Bias is indicated if the difference is large between the two. The 95-percent confidence level shown in Figure 7 should be analyzed prior to drawing any conclusions.

- Spatial bias--is similar to general bias, but differs in that it focuses on the spatial distribution. Mainly, we wish to determine if there is a spatial pattern to bias. While clever application of other tests could serve the purpose, the interstation error correlation test specifically addresses spatial bias. Also, the graphical output feature from the accuracy score tests provides a visual correlation for "spatial" performance.
- Temporal bias--the residual time series test is the only test specifically designed to provide a direct indicator of temporal (or cyclic) bias.
- Randomness--the bivariate regression and correlation test provides the best indicators. A low Pearson's correlation coefficient summarizes "randomness", and the scatterplot provides a visual indication.
- Input-parameter dependence--the multiple regression of error residuals test specifically assesses the effect of each input parameter using an SSPS procedure.[3] In addition, it ranks each parameter's relative contribution to the error residual. Other statistical tests will also be used for this purpose. However, these other tests must first be applied to subsets of the data base that are categorized by input parameters; then comparisons among categorized results are made. We discuss this further later in this section.

The above breakdown summarizes what is initially looked for in performance indicators.

Up to now we have been concentrating on application of the tests to the entire data base. The preceding paragraph discusses the general types of interpretation that lend themselves to our model evaluation process. In the next subsection, we describe what actions can be taken after this interpretation phase. First, however, we will look at each test individually and discuss how to interpret the results. (Refer to Volume 2 for a more comprehensive description.)

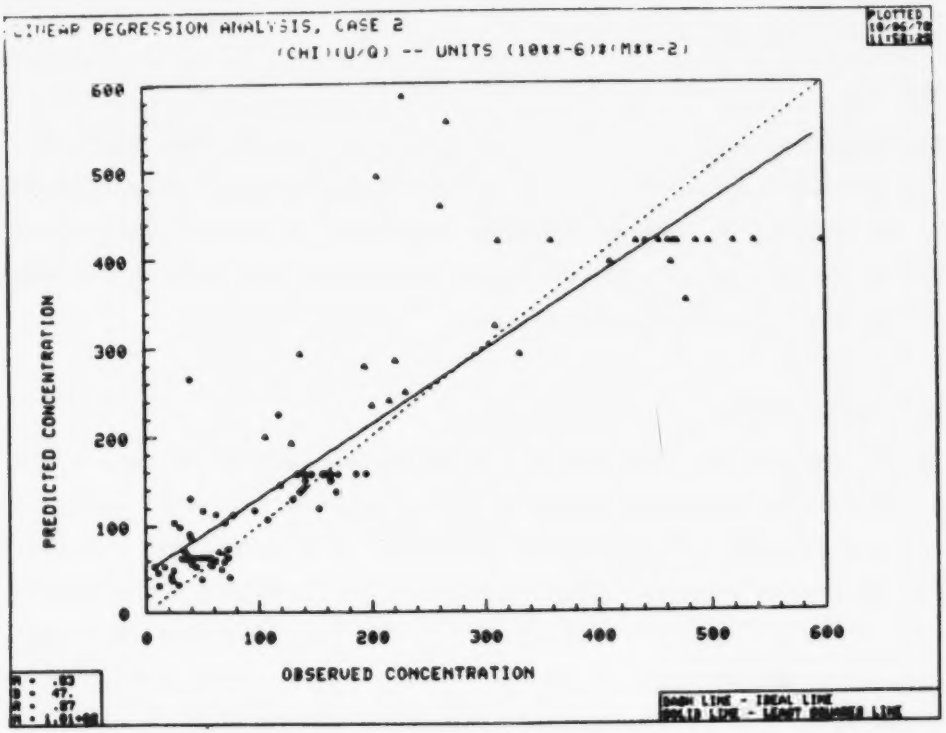



Figure 6. Actual versus ideal fit for test data base.

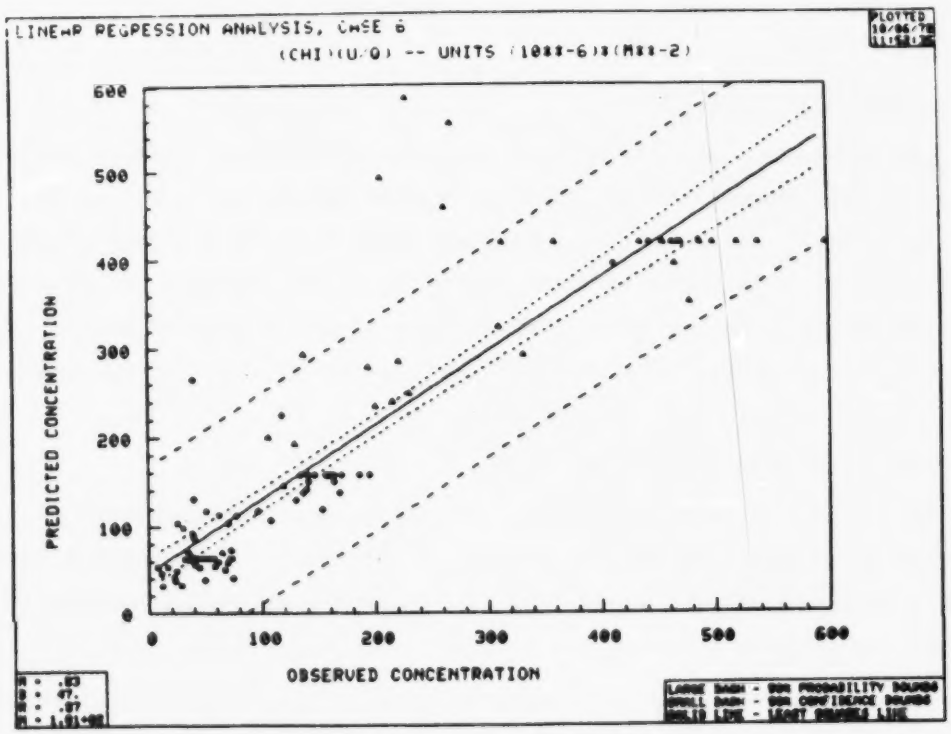



Figure 7. Scattergram with bounds for the test data base.

## Accuracy

The output of each accuracy score test is a number that becomes lower as agreement between observed and predicted concentrations becomes better. A confidence interval is also computed. A narrow range about the accuracy score is desired; the narrower the range, the more confidence we have in the accuracy score computation.

## Residual Time Series

Sample outputs for the residual time series test are given in Figures 8 and 9. The autocorrelation plot is examined to see if there is a large cyclic component in the error residual. The normalized cumulative periodogram yields much the same information but provides a cumulative display. The autocorrelation function is somewhat easier to interpret. Here one looks for the highest "spike" and, if it is above the 95-percent confidence line, one can assume that the model performance is significantly poorer for that time lag. In our example (Figure 8), we see that the autocorrelation is greatest for a time lag of 1.

## Chi-Square Goodness-of-Fit

The chi-square goodness-of-fit compares the distribution of observed and predicted concentrations (example outputs are given in Volume 3). As noted before, the logarithm-frequency plots provide the same graphical information in much the same form. The $\chi^2$ statistic summarizes the overall agreement of the two distributions in a single figure of merit. Like the accuracy score tests, a lower $\chi^2$ corresponds to a better agreement.

## Bivariate Regression and Correlation

The bivariate regression and correlation test is perhaps one of the most useful of our intermediate evaluation statistics. The output plots contain many useful visual displays to go along with useful numeric indicators. Individual features are described below:

- Pearson's correlation coefficient is an indicator of the linearity between predicted and observed concentrations. As such, it

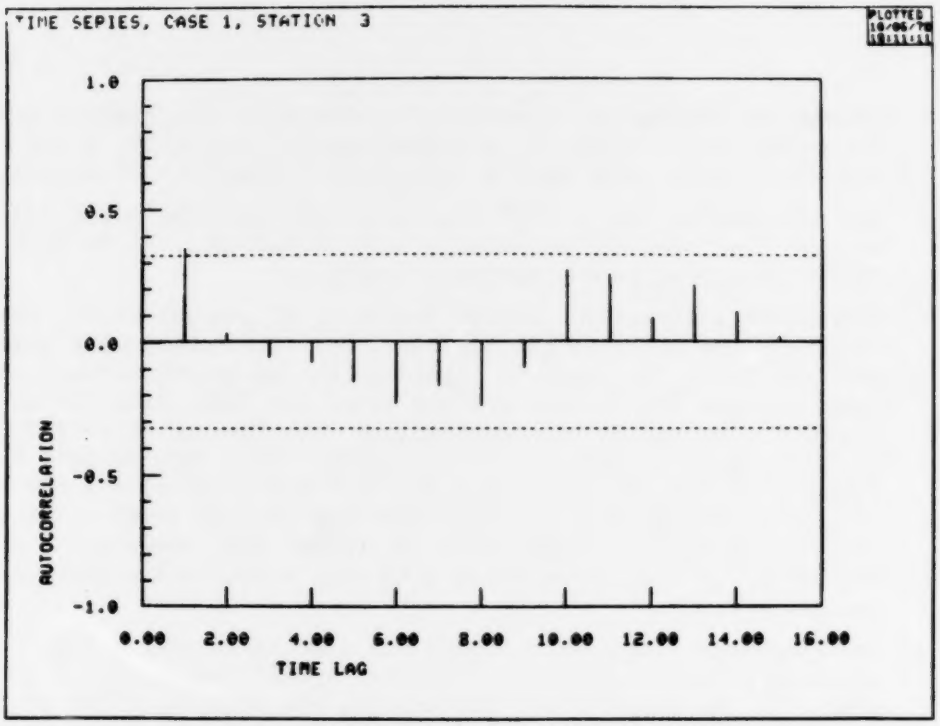



Figure 8. Autocorrelation function of test data base.

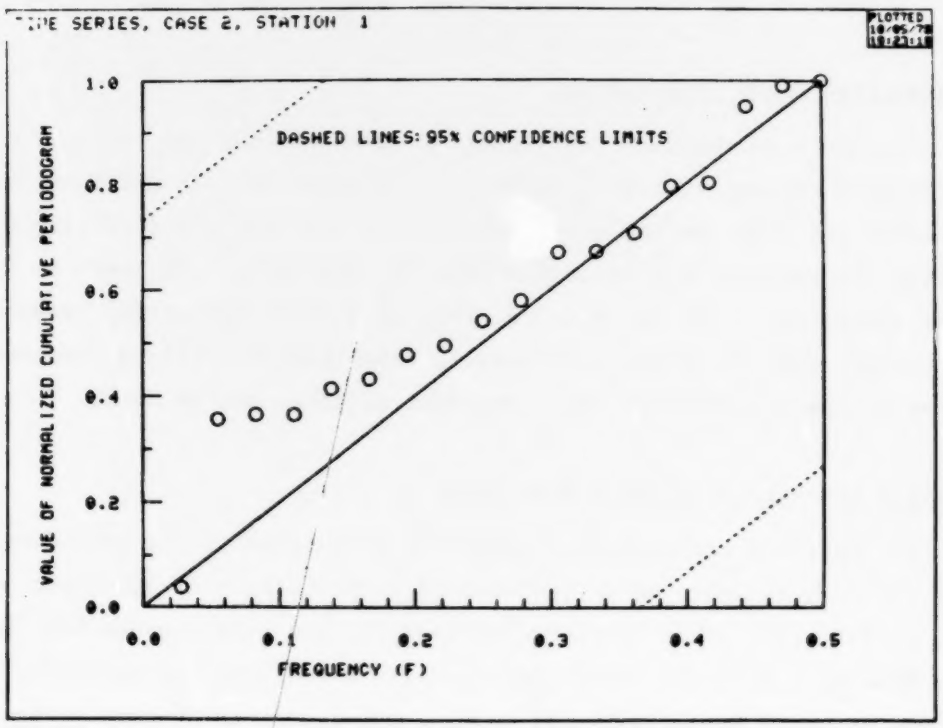



Figure 9. Normalized cumulative periodogram for test data base.

becomes an invaluable detector of randomness. The highest score, unity, basically rules out a random lack of agreement, while the low score, zero, indicates a completely random lack of agreement.

- The scatterplot is a visual display of the point-by-point comparison of predicted versus observed concentrations. It is best used with an auxiliary regression analysis.
- Regression analysis is another indicator of general bias. The slope and intercept of the least-squares line numerically quantifies the bias. As shown in Figure 6, the agreement between the least-squares (bias) line and the ideal (no bias) line can be visually compared on the scatterplot. The 90-percent confidence bounds (Figure 7) can also be displayed. These bounds tell the viewer that the real least-squares line will fall within the designated bounds at a statistically significant level. The 90-percent probability bounds tell the viewer that there are 9 chances out of 10 that a new point will fall within those designated bounds.
- User-supplied sensitivity bounds can also be entered. These sensitivity bounds essentially establish a dead zone on the scatterplot. In the subsequent calculations, the regression coefficients are based solely on the points falling outside the zone defined by dashed lines on the scatterplot. Thus, we can determine a trend in the so-called "outlying" comparison points.

Interstation Error Correlation

Pearson's correlation coefficient is computed for the residual concentrations between each pair of monitoring stations. If all correlations are about the same, we can probably conclude that the model is erroneously biasing the results in a uniform manner for all sites. If there is a large variation in the correlation, then we should examine the results for trends, such as differences between urban and rural sites, proximity to known local meteorological or emission factors, and so forth.

Multiple Regression of Error Residuals

The multiple regression of error residuals test is the only device to explicitly evaluate model performance as a function of the input parameters. (The other tests can be used for the same purpose but the data base must first be subdivided by parameter.) For the multiple regression test, essentially all model input parameters become independent

variables in the linear multiple regression equation; the concentration error residual is the dependent variable. The SPSS program computes the coefficients in the regression equation, the multiple correlation coefficient, and the contribution to this coefficient by each parameter. Hence, we have explicit indicators of which parameters contribute most to a model's poor performance.

Let us consider our example, in which the multiple regression test is applied to our test data base. The SPSS summary output for the test data base is given in Figure 10. There are three independent variables:

- Atmospheric-stability class (ASC)
- Wind-direction class (WDC)
- Wind-speed class (WSC).

The summary table illustrates the multiple correlation coefficient (MULTIPLE R) in the first column. The SPSS program arranged the parameters in the order of their contribution to MULTIPLE R. By far the most substantial contribution comes from the ASC parameter, with a value of 0.62443. Going down the list, the contribution from each of the other parameters is negligible because the MULTIPLE R increases only to 0.62586. For another way of looking at the results, note that the R SQUARE column shows that the first parameter, ASC, explains about 39 percent of the variance in the residuals; the other two, WDC and WSC, do not substantially account for more of the variance. The third and fourth columns, RSQ CHANGE and SIMPLE R, give the incremental change to the MULTIPLE R and the bivariate correlation coefficient, respectively.

The fifth column, B, gives the regression coefficients. Hence, the linear equation becomes:

$$\text{RESIDUAL} = 487.89 - 125.27\text{ASC} + 5.21\text{WDC} + 2.23\text{WDC} + \epsilon_1 \quad ,$$

where $\epsilon_1$ is the residual or the number required to equate both sides of the equation. However, the table shows that the contribution from the

```
TEST RUN NO. 1                                                  06 OCT 78        PAGE     6

FILE   SPSS:DATA (CREATION DATE = 04 OCT 78)

* * * * * * * * * * * * * * * * * * * * * * * * M U L T I P L E   R E G R E S S I O N * * * * * * * * * * * * *     VARIABLE LIST   1
                                                                                                                  REGRESSION LIST   1

DEPENDENT VARIABLE..    RESID      RESIDUAL

                                                SUMMARY TABLE

VARIABLE                          MULTIPLE R   R SQUARE   RSQ CHANGE   SIMPLE R          B              BETA

ASC                                 .62943      .39591      .39591      -.62943     -125.26599        -.63949
WDC                                 .62956      .39633      .00042      -.35899        5.20573         .04157
WSC                                 .62966      .39670      .00037       .32352        7.23051         .02257
(CONSTANT)                                                                            487.88769
```

Figure 10. SPSS summary output for test data base.

last two parameters is negliglble. Therefore, as a good approximation, our equation simplifies to:

$$\text{RESIDUAL} = 487.89 - 125.27\text{ASC} + \epsilon_2 \quad ,$$

where $\epsilon_2$ differs from $\epsilon_1$.

Basically, we are being told that improvement in the ASC parameter alone offers the potential for decreasing the variance in concentration residuals. Therefore, our approach would be to improve our scheme in estimating atmospheric stability (i.e., improve our model). On the other hand, it appears that our parameterization of wind speed and wind direction is all right. Therefore, little is to be gained from improving the model interface for wind speed and direction.

The summary table is only one part of the output resulting from the multiple regression test. The remaining output and further description of the SPSS results for the test data are given in Appendix A. In particular, we discuss statistical significance (the F statistic).

## DIAGNOSING THE CAUSES OF POOR MODEL PERFORMANCE

The example of the previous subsection illustrates results that were quite easy to interpret--namely, that atmospheric stability parameterization was a primary contributor to poor model performance. The results from the multiple regression tests will not always be so easy to interpret. Instead, they will be used to augment results from other tests to obtain a clear definition of the causes of poor model performance.

Our recommended approach in evaluating RAM is first to conduct any test that will provide useful insight into the problem prior to attempting any improvements to the RAM. The rationale is simple--in contrast to the model modification process, the statistical tests are inexpensive to run. After any model modification, it is necessary to generate a new test data base by running RAM using RAPS data for one year. This is a very expensive process which should take place only after we have obtained the maximum information from our diagnostic tests.

Diagnosis of the causes of poor model performance will normally be an iterative process, as shown in Figure 1 (Section 1). The iterative diagnostic process is illustrated as OPTION 3. As stated earlier in Section 2, we recommend application of all six tests to the entire data base as a first step in the diagnostic process. The results are then analyzed for our indicators of poor model performance: general bias, spatial bias, temporal bias, and randomness (see Section 2). Next, we must use our tests to relate these indicators to the causes of poor performance in terms of RAM parameters. There is no predefined prescription of how to proceed with our diagnosis. However, if we look at each of the indicators separately, there are certain recommendations that can be made on how to apply and interpret further tests.

In the following subsections, we address possible actions to explain and diagnose indicators of poor model performance.

General Bias

General bias exists when predicted and observed concentrations differ by a uniform amount (high or low) throughout the data base. This type of bias can easily be diagnosed by several of our tests, as described earlier. Using assumed displays from the bivariate regression and correlation test, let us consider various ways bias can manifest itself by comparing the least-squares regression line and the ideal fit. Figure 11 shows three different types of overprediction bias. Of the three, the second (diverging) is generally the most common in model evaluation. (This happens because the general wind direction is usually known and leads to a low concentration prediction to correspond with conditions when the monitor was upwind of the source area and recording low concentrations. For downwind conditions, the full diffusion equation comes into play and small errors in formulation or input parameters lead to large errors in concentration prediction.) The problem is, once we have detected a bias, what action do we take to explain it?

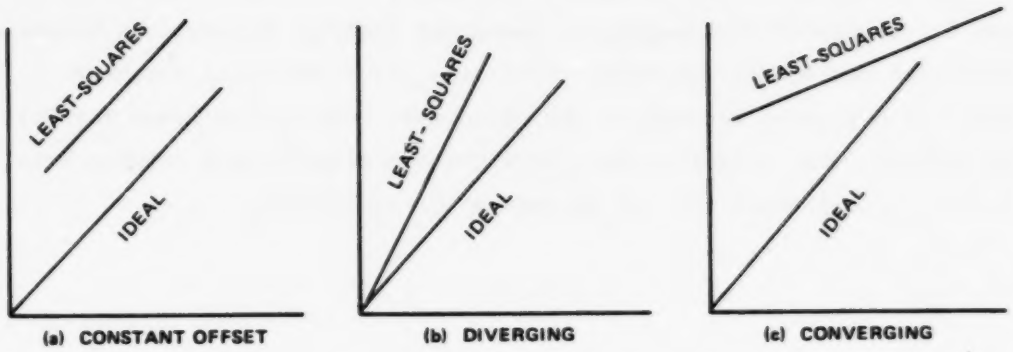



Figure 11. Types of bias.

At this point we must go into more detail. The first question concerns whether we are dealing with general bias or spatial bias, so the next step is to run the bivariate regression and analysis test for each monitoring site. If the comparisons between the ideal and least-squares lines are essentially the same for each station, then we have general bias. Next, we have to try to isolate the cause.

Results from other tests can help in diagnosing the causes of general bias. In particular, the correlation of error residuals among monitoring sites (interstation error correlation) can provide useful information. If all correlations are high, we are being told that the overprediction (or underprediction) generally exists for all stations at the same time. This not only confirms our diagnosis of general bias but also tells us that the causes of mispredictions are probably the same for all locations. Next, we examine the results of the multiple regression of error residuals test. This will tell us which RAM input parameter is most highly correlated with the error residuals. For example, we might find that the error residuals are inversely correlated to wind-speed class (WSC) or, as in our test data base, atmospheric stability class (ASC).

Unexplainable Biases--

At this point, let us depart from the above example and recognize one limitation of our diagnostic tests for finding correlations between RAM input parameters and error residuals. If a RAM input parameter causes a high bias throughout the data base, none of our tests may help us explain this. For instance, consider the relationship between predicted concentration (X) and emissions (Q) as follows:

$$X = kQ\ f(M) \quad ,$$

where f(M) is the meteorological function and k is a constant. If Q is biased high, then X will always be high if all other representations are accurate. If Q is 25 percent high, then a constant 25-percent offset would exist in our error residuals. Our tests are geared to correlate error residual changes (variance) to input parameters. For a constant 25-percent offset in the error residuals, there would be no change and hence no correlation.

It is unlikely that the above condition will exist in reality; emissions will most likely be biased differently according to the type of source. Hence, certain stations would exhibit a different bias from others because they would be influenced by different sources.

Second Iteration--

After the detection of general bias and analysis of other results, we will want to take a closer look at the dependence of error residuals on RAM input parameters. As stated, the multiple regression test provides us with some indication of input parameter dependence. For general bias, we will want to stratify the data base by classes of that parameter which influences the error residual most (as determined by the multiple regression of error residuals test).

Stratification of the data base can be accomplished by the STAT06*SPSSSORT program (see Volume 3). In our previous example, where the multiple regression test showed a high correlation between error

residuals and wind-speed class, we might choose to conduct our evaluations with the data base divided into low and high wind-speed categories. We would then apply our SPSSSORT program twice and generate two files, one for wind-speed classes of 1, 2, and 3, and one for classes 4, 5, and 6. We can then apply our bivariate regression and correlation tests to each file separately. If we find that one file shows good agreement while the other shows poor performance, we have identified problems in the formulation. Depending on how clear-cut the results are, we may wish to conduct further tests or proceed to consider modifications.

It must be recognized that there is no preset prescription for diagnosing model performance, and that our description here is for illustrative purposes only.

Spatial Bias

Probably the best initial test for spatial bias is the interstation error correlation test. As stated in Section 2, the output correlation matrix indicates the existence of a general bias if the correlation coefficients among all stations are high (say, 0.5 or above). Conversely, low correlation coefficients suggest the possibility of spatial bias. However, further test results are needed to confirm our suspicion because low correlation coefficients also result when any of the other indicators (e.g., randomness) of poor model performance exist.

The map display of the accuracy score test can also lend insight into spatial bias. If accuracy scores vary greatly from station to station, we again suspect spatial bias, but need to confirm our suspicion. Perhaps the best indicators for spatial bias are the bivariate correlation and regression displays applied to each station. The test for site-specific bias is then exactly the same as it was for general bias (see Section 2).

For the RAM evaluation, we must establish the degree of spatial bias for 13 stations. Next, our results must be categorized--namely, which stations can be thought of as being similar in performance? Referring

to our least-squares test versus the ideal fit (Figures 6 and 7), we can broadly categorize stations according to actual and ideal agreement. Results from the interstation error correlation test provide indicators of which stations behave similarly. Hence, we have the means for establishing reasonable categorizations of stations with similar performance.

Using the resulting categorizations, we look for rather obvious differences among locations, such as:

- Urban versus rural sites
- Eastern versus western sites
- Local influences (anomalous locations)
- Distribution in relation to large point-sources.

Furthermore, if we can group by similar stations, the next step is to use our sorting program to establish data bases for these similar stations. Then we apply the multiple regression of error residuals tests to these smaller data bases and proceed as we described in Section 2 (General Bias). It must be recognized, however, that each categorized data base may lead to different conclusions and subsequent remedial actions.

Before concluding our description, we caution that it is possible to have a situation in which certain stations exhibit spatial bias while others do not; other sites, for instance, could exhibit completely random but unbiased behavior.

Temporal Bias

The residual time series test is the only one capable of directly detecting temporal bias. Either the autocorrelation function or the normalized cumulative periodogram contains the same diagnostic information. Consider the example for the autocorrelation function A(k) in Figure 12. Temporal bias manifests itself when A(k) exceeds the 95-percent confidence limits. In Figure 12, temporal bias is shown to occur every 5 hours.

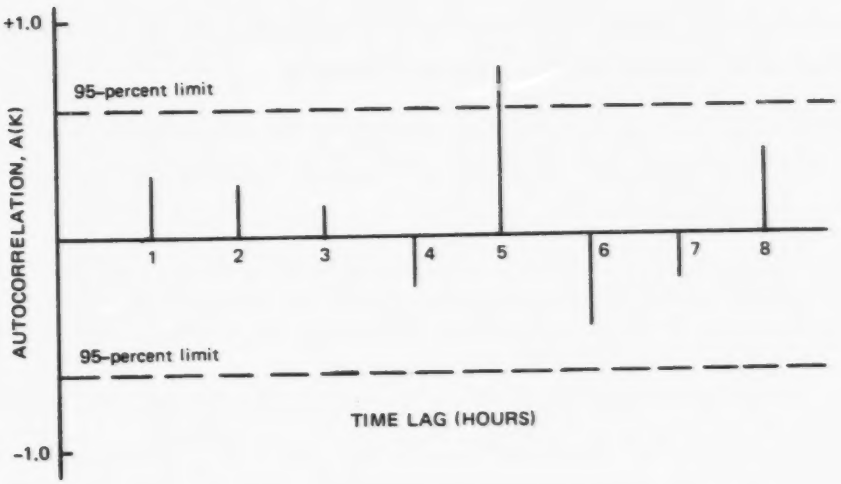



Figure 12. Temporal bias example.

Once detected, the cause of temporal bias must be diagnosed. We can look for the obvious, such as weekend versus weekday (lags of 6 or 7 days), day versus night (lags of 12 hours) or daily differences (lags of 1 day). When obvious conclusions cannot be reached, we suggest conducting tests for randomness.

Randomness

When randomness exists in the absence of bias (general or spatial), we would expect large error residuals but fairly good agreement between observed and predicted concentration distributions. Likewise, the least-squares regression line could compare quite favorably with the ideal line; however, the scatter by visual inspection would appear quite high and the correlation coefficient would be low (about 0.6 or less).

After randomness is detected, we suggest analyzing those point outside the 90-percent probability bounds (or subjectively selected sensitivity bounds) in conjunction with the results from our multiple regression of error residuals test, and with the results from the previous analysis of our data base (Section 2). Basically, we wish to determine which model input conditions exhibit the most influence on poor performance. Therefore, by comparing the input parameter distributions for

points outside the probability (or sensitivity) bounds with the input parameter distributions for the entire data base, we can determine which conditions most often lead to poor performance.

Let us consider the specific, step-by-step procedure along with an example from our test data base to illustrate the recommendations in the preceding paragraph:

Step 1: Conduct the bivariate regression and correlation test, plotting either probability or sensitivity bounds on the scatter-plot. For our example, consider those points outside the probability bounds (see Figure 7).

Step 2: Determine some logical criteria for specifying the majority of points falling outside the probability bounds. Referring to Figure 7, we see that most of the outlying points occur when predicted concentrations exceed 400 units* and observed concentrations are below 300 units.

Step 3: Modify the SPSSSORT program so that it displays the input parameters for those cases that meet the criteria established in Step 2. For application to our test data base, we added the following statements to our SPSSSORT program:

```
*SELECT IF        (PC GT 400 AND OC LT 300)
LIST CASES        VARIABLES=ALL
FREQUENCIES       INTEGER=WSC(1,6)
```

The first statement performs the sorting of the data base, given our Step 2 criteria. LIST CASES then lists all the parameters in the data base for those cases satisfying our criteria. The FREQUENCIES statement plots a histogram of the desired input parameter for the criteria case. A complete listing of program and output is given in Appendix B.

Step 4: Analyze the Step 3 results and compare the frequency distributions with those obtained earlier for the whole data base. Referring to our examples in Appendix B, we see that four cases met our Step 2 criteria. All of these were for an atmospheric-stability class (ASC) of 5--as shown on page 3 (Appendix B) of output. Further, on page 4, we see that all 4 cases were in the low wind-speed category. (Generally, we will be working with frequency distributions for other input parameters. However, for 4 cases we can easily review the case-by-case

---

*For our test data base, units are $m^{-2} * 10^{-6}$; for RAM, units will be either $\mu g\ m^{-3}$ or ppb.

listings.) If we examine the atmospheric stability categories (ASC) for the entire data base (as shown in Volume 3), we see that the most frequent ASC is 4. Therefore, our results suggest that our prediction mechanism is giving us the greatest error when an ASC of 5 is input.

Step 5: Refine the analysis for the input parameter to which the model is most sensitive. Basically, we would use the SPSSSORT program to stratify our data base by that parameter. For our test example, we would form a small data base for each ASC class. Then we would apply our bivariate regression and correlation test to each of these small data bases.

Step 6: Identify other sensitive input parameters using the above steps. We wish to determine if our model is mispredicting because of poor parameterizations for more than one input parameter.

The above procedure is by no means unique and will not answer all our questions about why the model is doing poorly in certain situations. However, in concert with other tests, we have enough diagnostic tools to quantify conditions leading to poor performance.

### Input-Parameter Dependence

As stated above, the multiple regression of error residuals is the only test that directly addresses input-parameter dependence. However, as shown in the previous four subsections, other tests in conjunction with SPSSSORT-stratified data bases can address the problem. The multiple regression test itself can be applied to the stratified data bases. The net result would then be a clearer indication of the second most important parameter contributing to poor performance, i.e., high residuals.

## COMBINATIONS OF POOR PERFORMANCE INDICATORS

Our previous discussion very neatly separated poor performance indicators into categories. Now, it must be recognized that there may be several indicators that must be treated. For instance, we could very easily have high general bias and randomness at the same time. As a further complication, the underlying causes may be the same or different.

No matter how complex the problem appears to be, our tests can be applied in the manner described in this text. As stated before, there is no set formula for applying a specific test at a certain point. Sometimes highly subjective decisions will enter into the process. Nevertheless, the tools of analysis remain the same and provide sufficient means for assisting in the diagnosis of poor model performance.

## REFERENCES


1. A. D. Busse, and J. R. Zimmerman, "User's Guide to the Climatological Dispersion Model," Publication EPA-R4-73-024, Environmental Protection Agency, Research Triangle Park, North Carolina (1973).

2. F. L. Ludwig et al., "A Practical Multipurpose Urban Diffusion Model for Carbon Monoxide," Final Report, SRI Project 7874, SRI International, Menlo Park, California (1970).

3. N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, D. H. Bent, Statistical Package for the Social Sciences (McGraw-Hill Book Co., New York, New York, 1975).

4. F. Smith and R. Strong, "Independent Performance Audit #4," Interim Report, EPA Contract 68-02-2407, Research Triangle Institute Project 43U-1291-1, Research Triangle Park, North Carolina (1977).

## APPENDIX A

## SPSS MULTIPLE REGRESSION OUTPUT FOR TEST DATA BASE

### SOURCE LISTING

Specification of the SPSS REGRESSION procedure is described in Volume III of this report. Referring to the first page of the program output (located at the end of this appendix, the source listing, note in particular the "(1)" in the field of the REGRESSION specification. This makes the program order the input parameters as a function of their significance or correlation with the error residual. Hence, the parameter entered first (by the program) is the most highly correlated variable; the second is the second most highly correlated variable, and so forth.

### UNIVARIATE STATISTICS

The program when "ALL" statistics are specified automatically computes some auxiliary statistics. This includes the mean, standard deviation, and number of cases, as illustrated on the second page of our test example.

### BIVARIATE CORRELATIONS

Bivariate correlation coefficients are also printed out. See the third page of the output listing.

### STEP-BY-STEP PARAMETER ENTRIES

The step-by-step entry of variables and associated statistics is illustrated on the fourth and firth pages of the sample output. In Table A-1, definitions of a number of the output statistics are presented. Of particular importance are the MULTIPLE R (as described in the test) and

TABLE A-1. DEFINITION OF SELECTED REGRESSION STATISTICS*

| Name | Description |
|---|---|
| MULTIPLE R | Multiple correlation coefficient |
| R SQUARE | Square of the correlation coefficient, which is normally equated to the fraction of variance of the dependent variable explained by the independent variables |
| ADJUSTED R SQUARE | R SQUARE corrected by an SPSS procedure according to the number of variables and cases |
| STANDARD ERROR | Standard error of estimate for the prediction of the residual concentration |
| DF | Degrees of freedom (or the number of independent variables) |
| RESIDUAL | Number of cases - DF - 1 |
| F | Significance indicator, which is the same as $t^2$ of student's t test. A table must be used to equate the F value to "statistical significance" or confidence levels. More formally, $F = \frac{\text{R SQUARE/DF}}{\text{(1 - R SQUARE)/RESIDUAL}}$ |
| B | Coefficient of the independent variable in the linear regression coefficient |
| BETA | Same as B except normalized to a unity value for the dependent variable |
| CONSTANT | Constant in the linear regression equation |

*Refer to Section 20 of the SPSS Manual[3] for comprehensive definitions of all parameters.

F, which is a measure of statistical significance. The reader is referred to the SPSS manual for a comprehensive description of all the statistics.

Examining the sample output, one can see that atmospheric stability class (ASC) is entered in the first step and has a 0.62443 correlation coefficient and a highly significant F value of 63.3. In step 2, wind-direction class (WDC) is entered but fails to improve the correlation coefficient appreciably while decreasing the significance (F = 31.5). On the third step (fifth page), wind-speed class (WSC) is entered with results even less significant than in the previous step.

SUMMARY TABLE

A summary table is printed on the last page of the output listing. The input parameters are listed in the order of their importance.

```
@SPSS
SPSS 6.04   SEPTEMBER 9, 1977

          SPACE ALLOCATION FOR THIS RUN..

              TOTAL AMOUNT REQUESTED                          9000 WORDS


              DEFAULT TRANSPACE ALLOCATION                    1125 WORDS


                 MAX NO OF TRANSFORMATIONS PERMITTED      37
                 MAX NO OF RECODE VALUES                  150
                 MAX NO OF ARITHM.OR LOG.OPERATIONS       300

              RESULTING WORKSPACE ALLOCATION                  7875 WORDS

            1.      RUN NAME       TEST RUN NO. 1
            2.      GET FILE       SPSSDATA


                                FILE SPSSDATA HAS  12 VARIABLES

                                THE SUBFILES ARE..

                                           NO OF
                                NAME       CASES

                                SPSSDATA     101

            3.      REGRESSION     VARIABLES = TIME,SITE,ASC,WSC,WDC,RESID/
            4.                     REGRESSION = RESID WITH ASC,WSC,WDC(1)/
            5.                     REGRESSION = RESID WITH TIME,SITE(1)/
            6.      STATISTICS     ALL

***** REGRESSION PROBLEM REQUIRES     216 WORDS WORKSPACE, NOT INCLUDING RESIDUALS *****
```

FILE SPSSDATA ( CREATION DATE = 04 OCT 78)

| VARIABLE | MEAN | STANDARD DEV | CASES |
|---|---|---|---|
| TIME | 1042.5743 | 469.3286 | 101 |
| SITE | 2.0000 | .8246 | 101 |
| ASC | 4.1782 | .3846 | 101 |
| WSC | 2.1683 | .7625 | 101 |
| WDC | 2.0891 | .6016 | 101 |
| RESID | -19.7391 | 75.3378 | 101 |

FILE SPSSDATA (CREATION DATE = 04 OCT 78)

CORRELATION COEFFICIENTS

A VALUE OF 99.00000 IS PRINTED
IF A COEFFICIENT CANNOT BE COMPUTED.

| | TIME | SITE | ASC | WSC | WDC | RESID |
|---|---|---|---|---|---|---|
| TIME | 1.00000 | -.02584 | .51154 | -.13348 | .58848 | -.26672 |
| SITE | -.02584 | 1.00000 | -.03153 | .03181 | .04031 | .02772 |
| ASC | .51154 | -.03153 | 1.00000 | -.47841 | .62214 | -.62443 |
| WSC | -.13348 | .03181 | -.47841 | 1.00000 | -.12022 | .32352 |
| WDC | .58848 | .04031 | .62214 | -.12022 | 1.00000 | -.35899 |
| RESID | -.26672 | .02772 | -.62443 | .32352 | -.35899 | 1.00000 |

FILE SPSSDATA (CREATION DATE = 04 OCT 78)

* * * * * * * * * * * * * * * * * * * * * * * * * * MULTIPLE REGRESSION * * * * * * * * * * * * * VARIABLE LIST 1
REGRESSION LIST 1

DEPENDENT VARIABLE.. RESID RESIDUAL

VARIABLE(S) ENTERED ON STEP NUMBER 1.. ASC

| | | | | | | |
|---|---|---|---|---|---|---|
| MULTIPLE R | .62443 | ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
| R SQUARE | .38981 | REGRESSION | 1. | 221304.39725 | 221304.39725 | 63.27114 |
| ADJUSTED R SQUARE | .38375 | RESIDUAL | 99. | 346273.78096 | 3497.71496 | |
| STANDARD ERROR | 59.14143 | | | | | |

VARIABLES IN THE EQUATION

| VARIABLE | B | BETA | STD ERROR B | F |
|---|---|---|---|---|
| ASC | -122.31519 | -.62443 | 15.37721 | 63.271 |
| (CONSTANT) | 431.27042 | | | |

VARIABLES NOT IN THE EQUATION

| VARIABLE | BETA IN | PARTIAL | TOLERANCE | F |
|---|---|---|---|---|
| WDC | .03214 | .03613 | .77112 | .128 |

* * * * * * * * * * * * * * * * * * * \

FILE SPSSDATA (CREATION DATE = 04 OCT 78)

* * * * * * * * * * * * * * * * * * * * * * * * MULTIPLE REGRESSION * * * * * * * * * * * * * VARIABLE LIST 1
REGRESSION LIST 1

DEPENDENT VARIABLE.. RESID RESIDUAL

VARIABLE(S) ENTERED ON STEP NUMBER 3.. WSC

| | | | | | | |
|---|---|---|---|---|---|---|
| MULTIPLE R | .62586 | ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
| R SQUARE | .39170 | REGRESSION | 3. | 222317.73245 | 74105.91082 | 20.81986 |
| ADJUSTED R SQUARE | .37288 | RESIDUAL | 97. | 345260.44576 | 3559.38604 | |
| STANDARD ERROR | 59.66059 | | | | | |

VARIABLES IN THE EQUATION

| VARIABLE | B | BETA | STD ERROR B | F |
|---|---|---|---|---|
| ASC | -125.26593 | -.63949 | 23.18490 | 29.191 |
| WDC | 5.20573 | .04157 | 13.10932 | .158 |
| WSC | 2.23051 | .02257 | 9.22273 | .058 |
| (CONSTANT) | 487.88763 | | | |

VARIABLES NOT IN THE EQUATION

| VARIABLE | BETA IN | PARTIAL | TOLERANCE | F |
|---|---|---|---|---|

MAXIMUM STEP REACHED

FILE SPSSDATA (CREATION DATE = 04 OCT 78)

• • • • • • • • • • • • • • • • • • • • • • • • • • • MULTIPLE REGRESSION • • • • • • • • • • • • • VARIABLE LIST 1
REGRESSION LIST 1

DEPENDENT VARIABLE.. RESID RESIDUAL

SUMMARY TABLE

| VARIABLE | MULTIPLE R | R SQUARE | RSQ CHANGE | SIMPLE R | B | BETA |
|---|---|---|---|---|---|---|
| ASC | .62443 | .38991 | .38991 | -.62443 | -125.26599 | -.63949 |
| WDC | .62556 | .39133 | .00142 | -.35899 | 5.20573 | .04157 |
| WSC | .62586 | .39170 | .00037 | .32352 | 7.23051 | .02257 |
| (CONSTANT) | | | | | | |

# APPENDIX B

## PARTIAL SAMPLE OF OUTPUT FOR SPSSSORT PROGRAM

```
@SPSS
SPSS 6.04   SEPTEMBER 3, 1977

          SPACE ALLOCATION FOR THIS RUN..

              TOTAL AMOUNT REQUESTED                          9000 WORDS


              DEFAULT TRANSPACE ALLOCATION                    1125 WORDS


                 MAX NO OF TRANSFORMATIONS PERMITTED        37
                 MAX NO OF RECODE VALUES                    150
                 MAX NO OF ARITHM.OR LOG.OPERATIONS         300

              RESULTING WORKSPACE ALLOCATION                  7875 WORDS

            1.      RUN NAME         SORTFILE
            2.      GET FILE         SPSSDATA


                              FILE SPSSDATA HAS  12 VARIABLES

                              THE SUBFILES ARE..

                                          NO OF
                              NAME        CASES

                              SPSSDATA     101


            3.      RAW OUTPUT UNIT 15
            4.      *SELECT IF       (ASC EQ 4)
            5.      WRITE CASES      (2F6.0,F4.0,2F3.0,F9.0,7X,2F8.1) DATE,TIME,SITE,ASC
            6.                       WSC,WDC,OC,PC
```

83 CASES WRITTEN ON LOGICAL UNIT #15
0 (UNWEIGHTED) CASES WERE DROPPED DUE TO MISSING VALUES

DATA TRANSFORMATION DONE UP TO THIS POINT..

| | |
|---|---|
| NO OF TRANSFORMATIONS | 2 |
| NO OF RECODE VALUES | 0 |
| NO OF ARITHM. OR LOG. OPERATIONS | 2 |

THE AMOUNT OF TRANSPACE REQUIRED IS 60 WORDS

```
   7.      *SELECT IF       (PC GT 400 AND OC LT 300)
   8.      LIST CASES       VARIABLES=ALL
   9.      FREQUENCIES      INTEGER= WSC(1,6)
  10.      OPTIONS          3,8
```

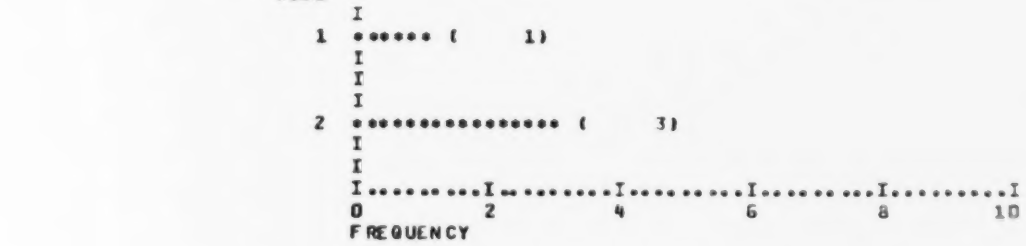
*FREQUENCIES* PROBLEM REQUIRES 37 WORDS OF SPACE

WSC

| CATEGORY LABEL | CODE | ABSOLUTE FREQUENCY | RELATIVE FREQUENCY (PERCENT) | ADJUSTED FREQUENCY (PERCENT) | CUMULATIVE ADJ FREQ (PERCENT) |
|---|---|---|---|---|---|
| | 1 | 1 | 25.0 | 25.0 | 25.0 |
| | 2 | 3 | 75.0 | 75.0 | 100.0 |
| | TOTAL | 4 | 100.0 | 100.0 | |

WSC

CODE




VALID CASES 4 MISSING CASES 0

FILE SPSSDATA (CREATION DATE = 30 NOV 78)

CONTENTS OF CASE NUMBER 1

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SEQNUM | 19. | SUBFILE | SPSSDA | CASWGT | 1.0000 | DATE | 741015. | TIME | 900. |
| SITE | 1. | ASC | 5. | WSC | 2. | WDC | 3. | OC | 230. |
| PC | 584. | RESID | -355. | | | | | | |

CONTENTS OF CASE NUMBER 2

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SEQNUM | 22. | SUBFILE | SPSSDA | CASWGT | 1.0000 | DATE | 741015. | TIME | 1000. |
| SITE | 1. | ASC | 5. | WSC | 2. | WDC | 3. | OC | 262. |
| PC | 457. | RESID | -195. | | | | | | |

CONTENTS OF CASE NUMBER 3

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SEQNUM | 24. | SUBFILE | SPSSDA | CASWGT | 1.0000 | DATE | 741015. | TIME | 1100. |
| SITE | 1. | ASC | 5. | WSC | 2. | WDC | 3. | OC | 259. |
| PC | 554. | RESID | -285. | | | | | | |

CONTENTS OF CASE NUMBER 4

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SEQNUM | 52. | SUBFILE | SPSSDA | CASWGT | 1.0000 | DATE | 741031. | TIME | 1800. |
| SITE | 1. | ASC | 5. | WSC | 1. | WDC | 3. | OC | 207. |
| PC | 491. | RESID | -284. | | | | | | |